Can You Trust "Black Box" AI?

Ensuring AI Safety for Healthcare

Using KERI to Secure Healthcare's AI

Executive Summary

- **Key Risks in Using AI for Healthcare**
  - While AI can enhance diagnostics, treatment, and efficiency it also introduces extreme risks related to data security, model reliability, and regulatory compliance. In particular; data integrity, attack vulnerabilities, lack of explainability, misinformation, and bias are major concerns.

- **An Open Source Solution**
  - KERI (Key Event Receipt Infrastructure) ensures verifiable data provenance, AI model integrity, zero-trust access, and real-time integrity checks.
  - KERI is adopted within the Trust Over IP (ToIP) Foundation and is referenced as part of an ISO Standard (*LEI ISO Spec, 17442*) through the Global Legal Entity Identifier Foundation (GLEIF).

- **Safe Data is Profitable Data**
  - Healthcare has the highest cost breaches/attacks of any industry.
  - Health systems must secure data before achieving data ROI

- **KERI + AI Use Cases**: Because KERI ensures data safety, a new horizon of AI-powered diagnostics, clinical decision support tools, research, and data sharing is now available.

There are critical risks associated with AI implementation in healthcare, this isn't news. However as the technology rapidly advances from novel to critical for solving healthcare's great problems, new solutions to the safety of PHI are required.

Here, we outline how KERI, an open-source, advanced protocol for key management, addresses the safety risks of AI by creating verifiable data provenance, immutable AI model lineage, zero-trust access controls, and real-time integrity checks.

healthKERI, Inc.
Jared Jeffery, FACHDM
Founder & CEO
info@healthkeri.com

Kerion, LLC
Samuel M. Smith, PhD
Founder & CEO
sam@kerion.one

Glossary of Terms

AI/ML (Artificial Intelligence / Machine Learning): AI refers to the simulation of human intelligence in machines, while ML is a subset of AI that enables computers to learn and improve from data without being explicitly programmed.

Black Box AI: A type of artificial intelligence system whose decision-making process is not easily interpretable by humans. These models, often based on deep learning, produce outputs without providing insight into how decisions were made, raising concerns in high-stakes fields like healthcare.

Shared Secrets: The traditional security method where authentication information (such as passwords or cryptographic keys) is known to multiple parties. This common, standard approach introduces vulnerabilities, as compromised secrets are the primary method hackers use to gain unauthorized access.

KERI (Key Event Receipt Infrastructure): An open source, cryptographic key management protocol that solves the hard problems of modern PKI (Private/Public Key Infrastructure) and is referenced as part of an ISO Standard. (*LEI ISO Spec, 17442*)

ISO (International Organization for Standardization): An independent, international body that develops and publishes standards to ensure quality, safety, and efficiency in various industries.

Data Safety: The practice of establishing trust in data by protecting it from corruption, unauthorized access, and breaches, ensuring reliability, transparency, and compliance with regulations.

Provenance: The ability to trace the origin, history, and modifications of data, ensuring authenticity and trustworthiness.

Zero-Trust: A security framework that requires continuous verification for access, assuming no entity is inherently trustworthy. "Never trust, always verify."

ACDC (Authentic Chained Data Containers): A cryptographically secure mechanism to ensure verifiable proof of data lineage, integrity, and provenance, crucial for AI and healthcare applications.

LEI (Legal Entity Identifier): A unique identifier assigned to legal entities participating in data transactions, ensuring transparency and regulatory compliance in global markets.

vLEI: A verifiable Legal Entity Identifier (vLEI) is a digital, cryptographically secure version of the LEI that extends the LEI's efficacy, enabling organizations to prove their identity in online transactions and regulatory reporting.

AI Risks in Healthcare

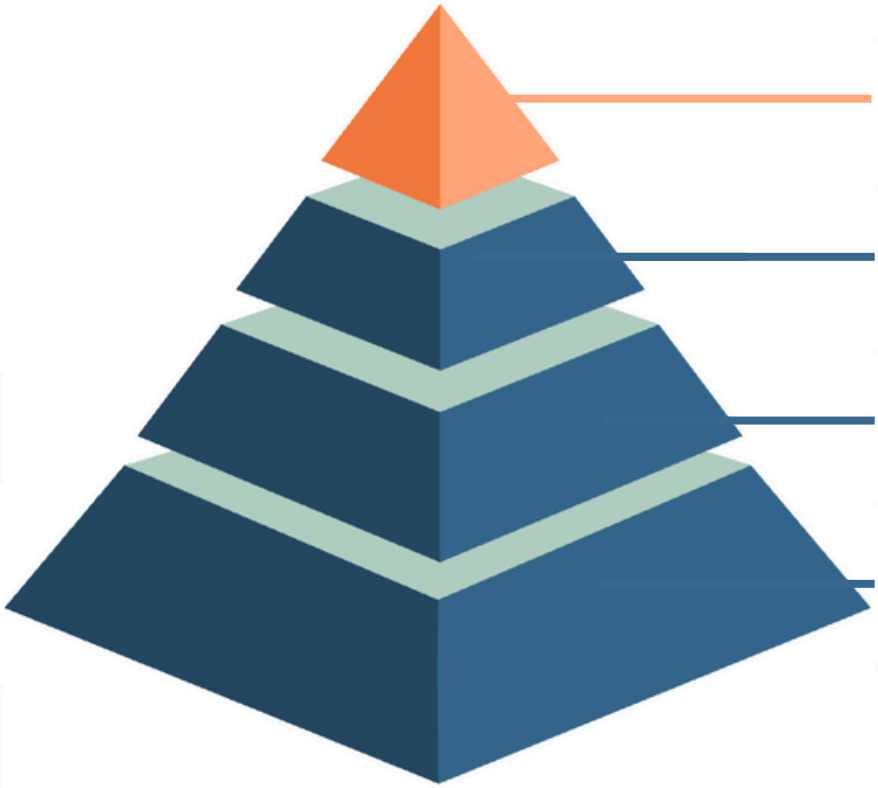
Data Integrity Risks
AI models rely on vast datasets, making them susceptible to data manipulation, poisoning attacks, or biased inputs, which can lead to incorrect predictions and poor patient outcomes. The MIT AI Risk Repository highlights concerns around AI reliability, emphasizing how adversarial attacks can subtly alter AI models to produce dangerous or misleading results.

Model Explainability & Accountability
Many AI models operate as “black boxes,” making it difficult to interpret their decision-making processes. This opacity is particularly dangerous in clinical decision-making where transparency is critical. MIT researchers have also emphasized the need for explainable AI in healthcare, advocating for techniques that ensure AI decisions can be traced and verified.

Security & Access Control
AI models must process sensitive patient data while ensuring compliance with HIPAA and other regulations. Traditional security approaches struggle to prevent unauthorized model manipulation or data breaches. The risks of AI-based cyber threats in healthcare have been widely documented, with reports showing that hospitals and medical institutions have increasingly become targets of ransomware attacks owing to the efficiency gains hackers achieve via AI.

Misinformation & Model Drift
In the modern iteration of an age old problem, AI suffers from "garbage in, garbage out." Models evolve over time, and sometimes in unintended ways. Without robust tracking and verification, an AI system can deliver outdated or misleading data that compromise safety. MIT has categorized AI misinformation as one of the top risks in its repository, citing examples where AI-generated predictions failed due to shifts in trends.

Incomplete Data Sets
AI systems trained on incomplete datasets—not reflective of a given population—can result in inaccurate outputs that result in suboptimal or dangerous outcomes. Studies have shown that AI-driven diagnostic tools can produce inaccurate outputs based on incomplete demographic data, leading to harmful healthcare outcomes (Harvard Medical School).



True zero-trust, next-generation data security

Authentic Chained Data Containers (ACDCs)
Creates verifiable data integrity

Key Event Receipt Infrastructure (KERI) - Solves the hard problems of decentralized key management.

Decentralized Public/Private Key Infrastructure (DPKI)
Used to sign / verify data

About KERI

KERI is an open-source, advanced protocol for key management that provides zero-trust, cryptographic assurances around data authenticity, security, and provenance. Outside of AI use cases, KERI has been adopted by the Trust Over IP (ToIP) Foundation, a global consortium focused on developing open governance and technology standards for verifiable digital trust ecosystems. Additionally, KERI has been recognized in the ISO standard for the LEI, as used by Global Legal Entity Identifier Foundation (GLEIF), an international organization that ensures reliable and standardized identity verification across industries.

Zero-Trust Access Controls

KERI eliminates the current reliance on shared secrets by enforcing continuous verification for data exchanges. This creates phish-proof data flows that prevent hackers from gaining unauthorized access.

Advanced PKI and Pre-Rotation

KERI operates by binding public/private key pairs to an Autonomic Identifier (AID), a self-certifying identifier that is cryptographically verifiable. Unlike traditional PKI-based systems, when initializing a KERI identifier, two cryptographic key sets are established: a current key set and a next key set.

The next key set is committed to within the current key state using a cryptographic digest. This ensures that future key rotations are predefined before they are activated, eliminating risks associated with key exposure and compromise.

By pre-rotating keys, KERI enforces forward security, so that even if the current key set is compromised, adversaries cannot manipulate future key states and recovery becomes as simple as a key rotation.

This pre-rotation also allows for "unbounded term AIDs", meaning that identities (organizational IDs, individual IDs) can persist indefinitely while maintaining cryptographic security through controlled, event-driven key rotations.

Core Components of KERI

Key Event Log (KEL)
A tamper-proof, append-only log that records cryptographic key events, ensuring a transparent and auditable chain of custody for data interactions.

Witness and Watcher Networks
Distributed networks of third party attesters who validate key events, preventing any single point failures and ensuring data authenticity.

Multi-Signature Functionality
KERI allows for multi-signature, threshold-based signing schemes to authorize key changes, enhancing security and resilience against cyber threats.

Event Anchoring
Via the KEL, KERI allows data to be 'anchored' into the keys state active at the time of anchoring, meaning that data can be provenanced with absolute assurance across time, even after keys have been rotated.

Leveraging these mechanisms, KERI creates an ecosystem where you never trust, but instead always verify both actors and their actions digitally. This significantly reduces the risks of unauthorized access and data manipulation of AI tools.

KERI and AI

AI inputs and outputs can be cryptographically signed and verified at the point of use, ensuring that recommendations have not been altered or manipulated. This capability is crucial in high-stakes environments such as medical diagnostics and clinical decision support.

This enables models and healthcare systems to maintain long-term, tamper-proof agents while seamlessly updating cryptographic credentials as needed.

Example Use Cases for AI + KERI in Healthcare

AI-Powered Diagnostics with Verifiable Trust
Hospitals implementing AI-based radiology interpretation can use KERI to ensure that AI models remain untampered and that every diagnostic output is cryptographically verified.

Clinical Decision Support Systems
AI-driven recommendations for treatment can be paired with KERI, ensuring that healthcare providers receive verified, auditable insights. This transparency increases their ability to trust the outputs of their AI.

Secure AI-Driven Research Data Sharing
Research institutions using AI can leverage KERI to guarantee the authenticity of shared datasets, mitigating risks of data corruption or misuse.

Regulatory Compliance & Audit Trails
AI model updates and decision-making processes can be cryptographically recorded using KERI, streamlining compliance with regulatory requirements.

AI Model Integrity & Supply Chain Security
KERI can be used to ensure the authenticity and integrity of AI models across their lifecycle, preventing unauthorized modifications or tampering. By cryptographically verifying AI model updates and ensuring a transparent lineage, KERI enhances trust in AI-powered applications, particularly in areas such as automated diagnostics, drug discovery, and healthcare operations.

Read the KERI Technical Whitepaper

Achieving the Returns on Investing in AI

Safe Data is Profitable Data

While AI represents an incredible opportunity for health systems to solve manual problems and improve patient outcomes—it also represents the next massive increase in data use. As the healthcare moves deeper into this digital transformation it will produce more data, more connections, and ultimately more risk.

The healthcare industry has become one of the most targeted sectors for cyber attacks, with incidents ranging from ransomware attacks on hospital networks to breaches of sensitive patient data. According to the Ponemon Institute, the average cost of a healthcare data breach in 2023 reached **$10.93 million** per incident, the highest of any industry. These costs include regulatory fines, legal fees, downtime, and reputational damage.

Operational Impacts of Unsafe Data

Beyond direct financial losses, cyber attacks severely disrupt hospital operations, delaying patient treatments and eroding trust in digital healthcare systems. A study by the American Hospital Association found that cyber incidents often result in weeks-long system downtimes, negatively impacting both patient safety and operational efficiency.

With healthcare data being a prime target for attackers, organizations must adopt a *true* zero-trust security framework to mitigate risks. KERI's cryptographic assurances represent an automated, scalable layer of security, ensuring data provenance, integrity, and secure access controls, reducing the likelihood of devastating cyber attacks.

Beyond mitigating risks, KERI enhances the return on investment (ROI) of AI implementations by ensuring that healthcare organizations can fully trust their AI-driven insights and comply with regulations without excessive security overhead. By reducing the risks of misinformation, data tampering, and compliance violations, KERI enables AI to operate more efficiently, securely, and with higher clinical and operational impact. Remember, because KERI is an open source technology, it can be utilized without vendor lock in.

Conclusion

AI Won't Fix Healthcare Until Healthcare Fixes Trust

As AI adoption in healthcare accelerates, organizations must prioritize security, trust, and data integrity. KERI offers a new and powerful, cryptographically backed approach to mitigating the risks posed by AI adoption. Through ensuring that healthcare's future AI systems remain transparent, trusted, and secure, KERI allows maximal ROI on data investments.

The adoption of AI in healthcare presents immense opportunities, but only if our industry implements it wisely. Healthcare leaders must take proactive steps to integrate data safety mechanisms, like those provided by KERI, to ensure AI operates with integrity. Waiting for regulations or relying on traditional security models is no longer a valid option. Next-generation tools demand next-generation security.

Sources

healthkeri.com/resources
https://www.gleif.org/en/vlei/introducing-the-verifiable-lei-vlei
https://airisk.mit.edu/
https://github.com/SmithSamuelM/Papers/blob/master/whitepapers/KERI_WP_2.x.web.pdf
https://www.eba.europa.eu/sites/default/files/2023-12/d5b13b4d-a9dc-4680-8b7c-0a3a4c694fac/Discussion%20paper%20on%20Pillar3%20data%20hub.pdf
https://www.ibm.com/reports/data-breach?

Secure Your Next AI Implementation